Acquiring Visual Classifiers from Human Imagination

Carl Vondrick, Hamed Pirsiavash, Aude Oliva, Antonio Torralba
Massachusetts Institute of Technology
{vondrick,hpirsiav,oliva,torralba}@mit.edu

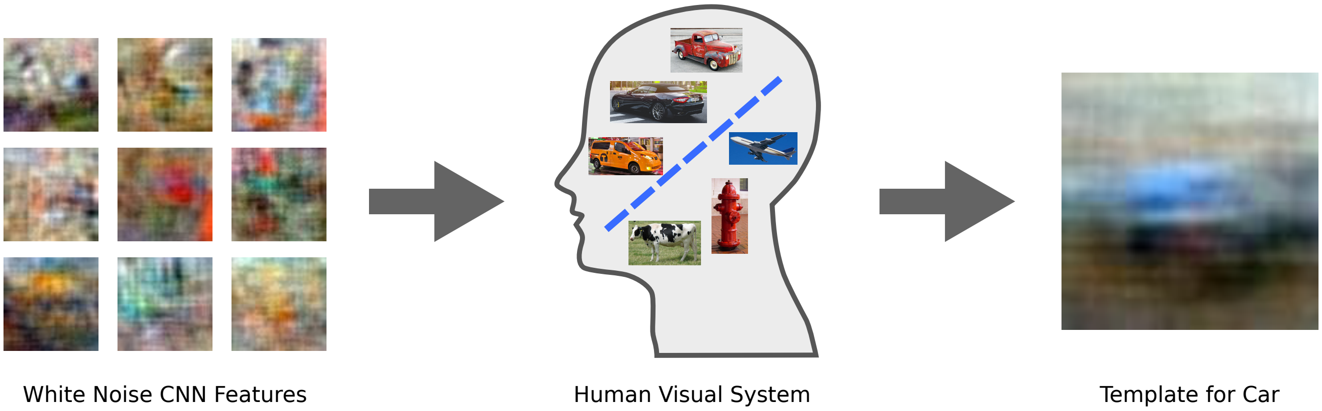


Figure 1: Although all image patches on the left are just noise, when we show thousands of them to online workers and ask them to find ones that look like cars, suddenly a car emerges in the average, shown on the right. This noise-driven method is based on well known tools in human psychophysics that estimates the decision boundary that the human visual system uses for recognition. In this paper, we explore how classifiers acquired from human imagination can be transferred into a machine.

Abstract

The human mind can remarkably imagine objects that it has never seen, touched, or heard, all in vivid detail. Motivated by the desire to harness this rich source of information from the human mind, this paper investigates how to extract classifiers from the human visual system and leverage them in a machine. We introduce a method that, inspired by well-known tools in human psychophysics, estimates the classifier that the human visual system might use for recognition, but in computer vision feature spaces. Our experiments are surprising, and suggest that classifiers from the human visual system can be transferred into a machine with some success. Since these classifiers seem to capture favorable biases in the human visual system, we present a novel SVM formulation that constrains the orientation of the SVM hyperplane to agree with the human visual system. Our results suggest that transferring this human bias into machines can help object recognition systems generalize across datasets. Moreover, we found that people's culture may subtly vary the objects that people imagine, which influences this bias. Overall, human imagination can be an interesting resource for future visual recognition systems.

1. Introduction

> *"Logic will get you from A to Z; imagination will get you everywhere."* — Albert Einstein

Computers routinely beat the human brain on challenges with logic and calculation speed. But, when it comes to object recognition, humans are still the state-of-the-art. What is the key difference between human recognition and machine recognition?

One answer is that the best object recognition systems today are unable to imagine objects that they have never encountered. However, the human mind can effortlessly imagine objects that it has never seen, touched, or heard. Even more remarkably, humans can do this in any color, orientation, deformation, put upside down, in and out of context, all in vivid detail.

In this paper, we seek to transfer the mental images of what a human can imagine into an object recognition system. We combine the strengths of two approaches: state-of-the-art features in computer vision [7, 23] with a method in human psychophysics [2] that estimates the decision boundary that the human visual system uses for recognition.

Consider what may seem like an odd experiment: we sample white noise in a visual feature space from a standard normal distribution. What is the chance that this sample is a car? Fig.1a visualizes some samples using feature inversion [38] and, as expected, we see noise. But, let us not stop there. We next generate one hundred thousand points from the same distribution, and ask workers on Amazon Mechanical Turk to classify each sample as a car or not. Fig.1c shows the average of visual features that workers believed were cars. Although our dataset consists of only white noise, a car emerges!

While sampling noise may seem unusual to computer

Report Documentation Page	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE **2014**	2. REPORT TYPE	3. DATES COVERED **00-00-2014 to 00-00-2014**

4. TITLE AND SUBTITLE **Acquiring Visual Classifiers from Human Imagination**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Massachusetts Institute of Technology,77 Massachusetts Avenue,Cambridge,MA,02139**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S)
	11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The human mind can remarkably imagine objects that it has never seen, touched, or heard, all in vivid detail. Motivated by the desire to harness this rich source of information from the human mind, this paper investigates how to extract classifiers from the human visual system and leverage them in a machine. We introduce a method that, inspired by wellknown tools in human psychophysics, estimates the classifier that the human visual system might use for recognition but in computer vision feature spaces. Our experiments are surprising, and suggest that classifiers from the human visual system can be transferred into a machine with some success. Since these classifiers seem to capture favorable biases in the human visual system, we present a novel SVM formulation that constrains the orientation of the SVM hyperplane to agree with the human visual system. Our results suggest that transferring this human bias into machines can help object recognition systems generalize across datasets. Moreover, we found that people?s culture may subtly vary the objects that people imagine, which influences this bias. Overall, human imagination can be an interesting resource for future visual recognition systems.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**10**	

vision researchers, a similar procedure, named *classification images*, has gained popularity in human psychophysics [2] for estimating the template the human visual system internally uses for recognition [20, 3]. In the procedure, an observer looks at random noise and indicates whether they perceive a target category. After a large number of trials, psychophysics researchers can apply basic statistics to extract the internal template the observer used for recognition. We discovered that a similar approach can be used to build a coarse object recognition system that originated from people's imagination.

Motivated by the observation that human visual system is a rich source of information, this paper investigates the scientific question whether visual classifiers acquired from human imagination can be leveraged computationally. Inspired by classification images, we introduce a method to estimate *imaginary classifiers* from the human mind, but in a feature space that is compact and discriminative for computers. To our knowledge, we are the first to extract classifiers from the human visual system in computer vision feature spaces. We then present a novel SVM formulation that integrates knowledge from the human visual system by constraining the SVM solution to be oriented close to the imaginary classifier. Our experiments are surprising, and suggest that classifiers from the human mind might be transferrable into a machine.

In addition, we found that imaginary classifiers are useful in two particular computer vision applications. Firstly, since these classifiers do not depend on real images, we can build recognition systems in situations where it is difficult to collect data. Our results suggest that it is possible to recognize objects in images in the wild without training on any real images. Secondly, since imaginary classifiers are estimated only by humans looking at noise, they inherit biases from the human visual system. Our experiments suggest that the bias from the human visual system is favorable, and can improve generalization performance across datasets. Overall, these experiments hint that human imagination can be an interesting resource for future visual recognition systems.

1.1. Related Work

This paper acquires a recognition system from the human mind by combining several popular methods. While each individual method is standard, their combination is novel. We briefly review the related work in both human and computer vision.

Mental Images: Our methods build upon work to extract mental images from a user's head for both general objects [16] and faces [26]. However, our work differs because we estimate mental images in state-of-the-art computer vision feature spaces, which allows us to integrate the mental images into an object recognition system.

Human-in-the-Loop: The idea to transfer classifiers from the human mind into object recognition is inspired by many recent works that puts a human in the computer vision loop [5, 10, 29], trains recognition systems with active learning [36, 34], and studies crowdsourcing [37, 32, 40]. The primary difference of these approaches and our work is, rather than using crowds as a workforce, we want to extract classifiers from the worker's minds using methods rooted in human psychophysics.

Transfer Learning: We build upon methods in transfer learning to incorporate priors into learning algorithms. A common transfer learning method for SVMs is to change the max-margin regularization term $||w||_2^2$ to $||w - c||_2^2$ where c is the prior [31]. However, this imposes an prior on the norm of of w. In our case, since the imaginary classifier does not provide an additional prior on the norm, we introduce a novel SVM formulation that constrains only the orientation of w to be close to c. Our approach extends sign constraints on SVMs [12], but instead enforces orientation constraints.

Deep Learning: There is a large body of work studying deep learning [24, 23], which hopes to build models that mimic neuron activations in the human brain. While our work is also inspired by biological vision, we are only interested in estimating the classifier parameters for very specific recognition tasks.

Human Psychophysics: Finally, our ideas extend classification images [20, 3], a tool in psychophysics to estimate decision boundaries that the human visual system uses. Firstly, while classification images have been mostly restricted to images and audio, we are the first, to our knowledge, to apply it to feature spaces in computer vision. Secondly, our approach uses only noise to estimate classifiers. Unlike classification images, we do not use any real images. We capitalize on the ability of people to discern visual objects from random noise in a systematic manner [17].

2. Acquiring Classifiers

In this section, we describe how to acquire classifiers from the human visual system. We first review a popular method in human psychophysics for performing this task. Then, we adopt it for use in a computer.

2.1. Classification Images

We first review *classification images*, a popular method in human psychophysics that estimates the internal template that the human visual system uses for recognition [20, 3]. The goal is to approximate the template $c \in \mathbb{R}^d$ that the human visual system uses for recognition.

The intuition behind classification images is simple, but powerful. We wish to discover how a human observer discriminates between two classes A and B, e.g. male vs. female faces, or chair vs. not chair. Suppose we have real images $a \in A \subseteq \mathbb{R}^d$ and $b \in B \subseteq \mathbb{R}^d$. If we sample white noise $\epsilon \sim \mathcal{N}(0^d, I_d)$ and ask an observer to indicate the class label for $a + \epsilon$, most of the time the observer will answer class A. However, there is a chance that ϵ might manipulate a to cause the observer to mistakenly label $a + \epsilon$ as class B.

The key insight into classification images is that, if we perform a large number of trials, then we can estimate a decision function $f(\cdot)$ that discriminates between A and B, but makes the same mistakes as the observer. Since $f(\cdot)$ makes the same errors, it provides a good model for the internal decision function that the observer uses. By analyzing this model, we can then gain insight into how the human visual system discriminates between A and B.

If we assume that the human visual system uses the linear decision boundary of the form $f(x; c) = c^T x$, then [27] shows that the classification image c with the optimal signal-to-noise ratio is:

$$c = (\mu_{AA} + \mu_{BA}) - (\mu_{AB} + \mu_{BB}) \tag{1}$$

where $\mu_{PQ} \in \mathbb{R}^d$ is the average image where the original was class P but the observer predicted Q.

Is it reasonable to assume that classification images should be linear? Although there is overwhelming evidence that object recognition in the human brain is nonlinear, a linear classification image is reasonable because we only seek an approximation of the human decision boundary. Moreover, while nonlinear models are possible, they require significantly more trials to estimate, which is expensive, and in practice we see good results with linear models. We do, however, wish to point out promising efforts that study nonlinear classification images [28].

2.2. Imaginary Classifiers

Since psychophysics researchers are interested in understanding how the human brain functions, they want to extract classifiers from the human visual system that are interpretable. Consequently, they build classification images in pixel space for geometric shapes or faces. However, we are interested in extracting classifiers to use in a computer. Inspired by classification images, we present an approach that acquires classifiers from the human visual system, but in the same feature spaces as computer vision systems. Our new method, which we refer to as *imaginary classifiers*, uses two key ideas.

Firstly, we captialize on recent work in feature inversion [38, 39, 8, 21]. Rather than generating noise in pixel space, we generate noise in feature space. We then invert the noise features back to an image and ask humans to label the feature visualization. Since machines understand features and humans understand visualizations, we are able to build a classifier that makes similar recognition mistakes as humans, but in a space that is discriminative for computers.

Secondly, we found that humans are surprisingly good at imagining objects in visualizations of feature space noise. When we instruct people to label visualizations of just white Gaussian noise (with no real images), people frequently find white noise that looks like objects. Feature descriptors often have structure (e.g., encodings of gradients or colors) that likely causes white noise in feature space to invert to images that look like objects. Although people are incorrect when they label pure noise as an object, they are providing information about how the human visual system discriminates objects in computer vision feature spaces.

We propose to build imaginary classifiers by combining feature inversion with people's ability to discern objects in pure noise. We first sample noise from a zero-mean, unit-covariance Gaussian distribution $x \sim \mathcal{N}(0_d, I_d)$. We then invert the noise feature x back to an image $\phi^{-1}(x)$ where $\phi^{-1}(\cdot)$ is the feature inverse. By instructing people to indicate whether a visualization of noise is a target category or not, we can build a linear classifier $\tilde{c} \in \mathbb{R}^d$ that approximates the decisions of their visual system:

$$\tilde{c} = \mu_A - \mu_B \tag{2}$$

where $\mu_A \in \mathbb{R}^d$ is the average, in feature space, of white noise that workers incorrectly believed was an actual object, and similarly $\mu_B \in \mathbb{R}^d$ is the average of noise that workers correctly labeled as noise. Since we sample white Gaussian noise, Eqn.2 can be interpretted as an LDA classifier [18] over labeled noise where the covariance is identity, $\Sigma = I$.[1] Moreover, observe Eqn.2 is a special case of the original human psychophysics Eqn.1 where the background class B is white noise and the positive class A is empty. Instead, we rely on humans to hallucinate objects in noise to form μ_A.

Since we average noise in feature space instead of pixel space, we have two advantages over standard classification images. Firstly, imaginary classifiers are in a feature space that is compact and discriminative, which allows us to plug the classifier into a machine. Secondly, since we build imaginary classifiers with only white Gaussian noise and no real images, our approach is immune to many issues in dataset bias [35]. Instead, imaginary classifiers inherit the biases present in the human visual system, which we suspect provides advantangeous signals about the visual world.

We were able to estimate $\tilde{c}$ with one hundred thousand trials. We picked an aspect ratio appropriate for our target category, sampled one hundred thousand points in feature space from the standard normal multivariate distribution, and inverted each sample with HOGgles [38]. We then put each visualization on Amazon Mechanical Turk [32] and instructed workers to indicate whether they see the target category or not. Since we found that the interpretation

[1] We tried training other classifiers too (such as SVM), but we did not see any advantage in our experiments.

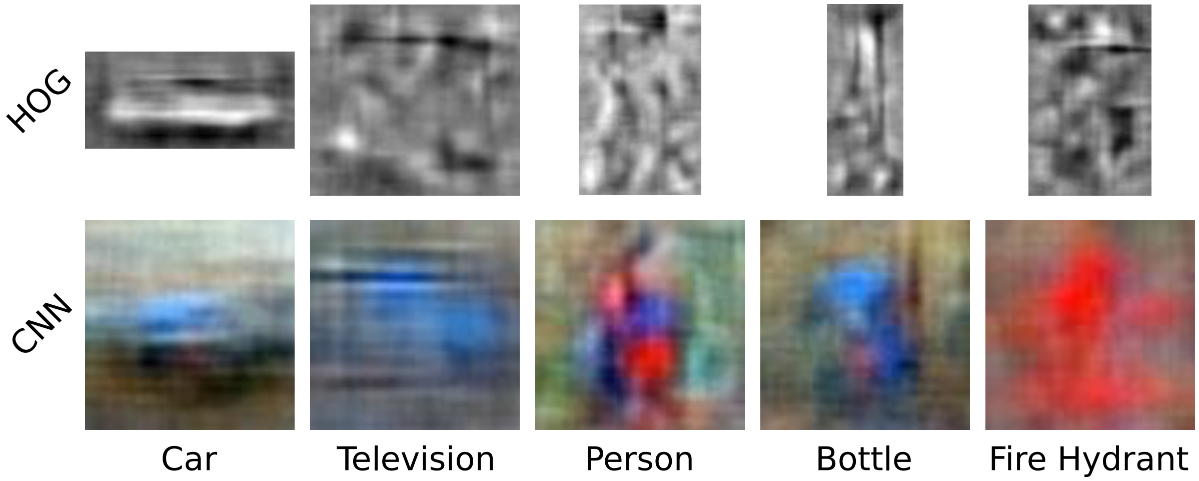


Figure 2: We visualize some decision boundaries acquired from the Mechanical Turk workers' minds. Although they are blurred, in many cases significant detail can be observed. Notice that the car classifier captures a darker road below the car, and a lighter sky towards the top. The television shows a rectangular structure, the person mimics a pedestrian, and the valves can be seen in the fire hydrant.

of noise visualizations depends on the scale, we show the worker three different scales. We paid workers 10¢ to label 100 images, and the workers were fast, often solving the entire one hundred thousand images in a few hours. Our experiments were affordable, with each classifier only costing around $100 to build. In order to assure quality, we occasionally gave workers an easy example to which we knew the answer, and only retained work from workers who performed well above chance. We only used the easy examples to qualify workers, and discarded them for computing the final classifier.

Surprisingly, although subjects are classifying zero-mean, identity covariance white Gaussian noise, objects suddenly emerge after many trials. We visualize some of the imaginary classifiers in Fig.2. In many cases, we can observe significant detail. For example, in the car classifier, we can clearly see a vehicle-like object in the center sitting on top of a dark road and light sky. The television clearly shows a rectangular structure, and the fire hydrant reveals a red hydrant with two arms on the side.

3. Experiments on Imaginary Classifiers

There is a large class of visual objects that humans can imagine, but they have never seen. In order to explore the extent that human imagination can play a role in computer vision, we want to scientifically understand how well we can acquire classifiers from the human visual system and leverage them computationally. Hence, we will evaluate how well we can extract imaginary classifiers by quantifying their ability to discriminate and recognize objects.

3.1. Experimental Setup

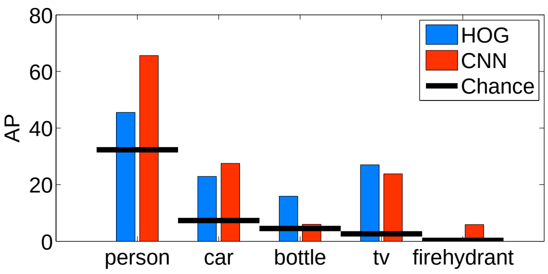
We evaluate our methods on object classification. We assume object localization is given and the task is to predict the category of each window. We conduct our experiments on the PASCAL VOC 2011 dataset [13], evaluating against the validation set.[2] We report performance as the average precision on a precision-recall curve. We show results for two sets of features: HOG [7] and the last convolutional layer (pool5) of a convolutional neural network (CNN) trained on ImageNet [23, 9]. We use the Felzenszwalb et al. implementation of HOG [15] and Decaf for extracting CNN features [11]. We trained inversions for both features with paired dictionary learning [38]. All classification images are estimated on Amazon Mechanical Turk with $150,000$ trials.

3.2. Evaluation

The results in Fig.3 suggest that our imaginary classifiers are capturing some signal from the human visual system. Although the classifiers are estimated using only white noise, in nearly every case the imginary classifiers are significantly outperforming chance. The delta in AP is occasionally large, with performance on person doubling and television performance increasing an order of magnitude.[3] These results suggest that we are able to acquire some signal from the human visual system and start to leverage it

[2] We added 63 annotated fire hydrants to the dataset for reasons that will become clear later.

[3] Although most researchers walk past a fire hydrant every day, they are not annotated in any major recognition dataset, including ImageNet. However, since imaginary classifiers do not require datasets, we are still able to recognize them!



	car	person	f-hydrant	bottle	tv
HOG	22.9	45.5	0.8	15.9	27.0
CNN	27.5	65.6	5.9	6.0	23.8
Chance	7.3	32.3	.3	4.5	2.6

Figure 3: We show the average precision (AP) for object classification on PASCAL VOC 2011 using the classification image. Even though the classification image was created without a dataset, it performs significantly above chance in nearly every case (green). If a machine learning algorithm were trained without data, the best it could do is chance.

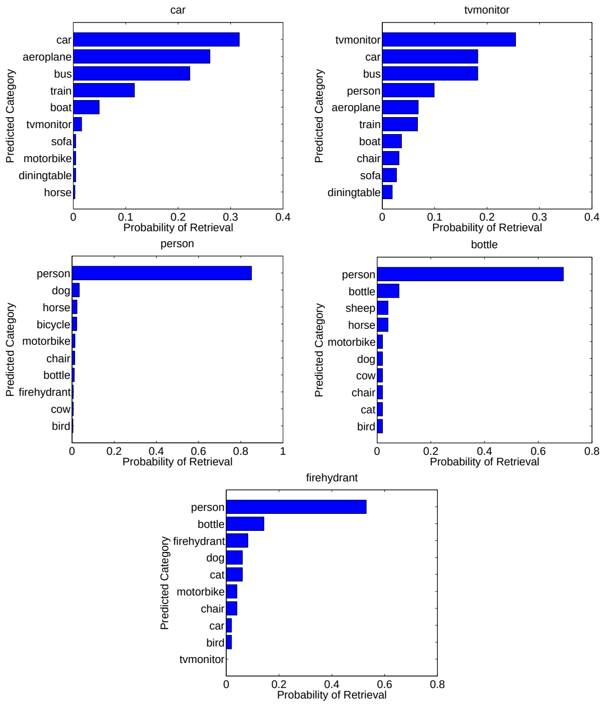


Figure 4: We plot the class confusions for each imaginary classifier for top classifications for CNN features. We show only the top 10 classes for visualization. Notice that many of the confusions are semantically meaningful, e.g. the classifier for car tends to retrieve vehicles, and the fire hydrant classifier commonly mistakes people and bottles.

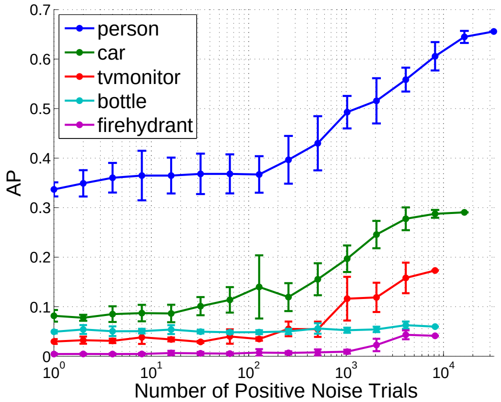


Figure 5: We plot object classification performance on PASCAL VOC with CNN features versus number of positive noise trials. As the number of trials increase, performance increases as well. Our results suggest that performance has not yet saturated for many categories. Error bars show standard deviation over 10 random samples.

computationally.

Moreover, the misclassifications for the imaginary classifiers are often sensible. Fig.4 shows the class confusions for the top classification for each classifier. Notice that cars are frequently confused with other vehicles, and bottles are commonly confused with people. We hypothesize that future work in building higher resolution classifiers will resolve some of these issues. The number of noise trials needed to estimate a imaginary classifier is also feasible, making the method affordable. Fig.5 shows performance versus the number of noise trials for a few categories. In many cases, 10,000 positive trials is enough to estimate a classifier. Performance does not appear to have yet saturated, suggesting that better classifiers can be created with more trials.

We note that one potential concern in our experiments is that the CNN features are trained to discriminate on ImageNet [9] LSVRC 2012, and hence had access to data. To address this concern, we have shown results for HOG as well, which is a hand-crafted feature. Additionally, we showed results for categories that the CNN network did not see during training (people and fire hydrants).

3.3. Analysis

Our experiments indicate that imaginary classifiers contain some discriminative power. By analyzing what the classifier is using for discrimination, we can gain insight into how the human visual system recognizes objects in computer vision feature spaces.

Our results suggest that shape is important for the imaginary classifier to discriminate in CNN feature space. Notice how the top classifications in Fig.6 tend to share the same

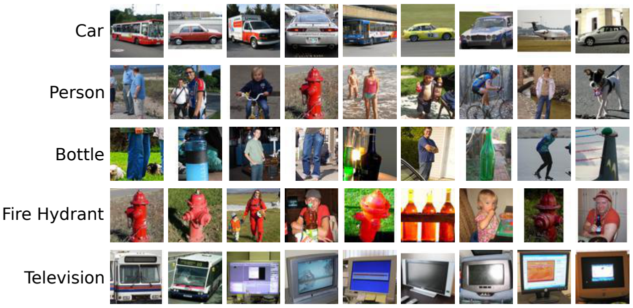


Figure 6: We show some of the top classifications from the imaginary classifiers estimated with CNN features.

rough shape by category. For example, the classifier for person finds people that are upright, and the television classifier fires on rectangular shapes. The confusions in Fig.4 confirm these findings since bottles are often confused as people, and cars are confused as buses. Moreover, the visualization of the classifers in Fig.2 attempts to show the canonical shape that the classifier has learned. Although the visualization is blurry, oftentimes one can see strong shape details, such as the valves appearing in the fire-hydrant. Indeed, shape seems to be important.

In addition to shape, some imaginary classifiers appear to rely on color as well. Fig.6 suggests that the classifier for fire-hydrant correctly favors red objects, which is evidenced by it frequently firing on people wearing red clothes. The bottle classifier, although, seems to be incorrectly biased towards blue objects, which contributes to its poor performance. We suspect that the Mechanical Turk workers likely subconsciously biased the bottle classifier towards blue. While color is not as important as shape, color appears to be useful for humans to recognize objects in noise.

These results suggest together that the human visual system encodes some bias towards the shape and color of objects. Since humans are the best object recognition agents, we suspect that this bias is favorable. In the remainder of this paper, we will explore how we can use these biases from the human visual system.

4. Learning with Imaginary Classifiers

Our experiments suggest that imaginary classifiers provide a good template for the features that the human visual system finds discriminative for recognition between two classes. Since the human visual system is the best object recognition system today, we suspect that integrating imaginary classifiers with machine learning can be powerful. In this section, we present a novel SVM formulation that incorporates knowledge from the human visual system by constraining the SVM hyperplane to have a similar orientation to the imaginary classifier.

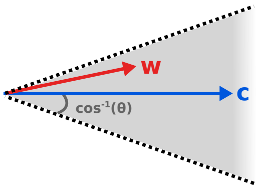


Figure 7: We visualize the conic constraint on the SVM solution w. The feasible space for the solution is the grayed hypercone. The SVM solution w is not allowed to deviate from $\tilde{c}$ by more than $\cos^{-1}(\theta)$ degrees.

4.1. SVM with Orientation Priors

Let $x_i \in \mathbb{R}^m$ be a training point and $y_i \in \{-1, 1\}$ be its label for $1 \leq i \leq n$. The SVM seeks a separating hyperplane $w \in \mathbb{R}^m$ with a bias $b \in \mathbb{R}$ that maximizes the margin between positive and negative examples. We wish to add the constraint that the SVM hyperplane w must be at least $\cos^{-1}(\theta)$ degrees away from the imaginary classifier $\tilde{c}$:

$$\min_{w,b,\xi} \frac{\lambda}{2} w^T w + \sum_{i=1}^{n} \xi_i \tag{3a}$$

$$\text{s.t.} \quad y_i \left(w^T x_i + b \right) \geq 1 - \xi_i, \quad \xi_i \geq 0 \tag{3b}$$

$$\theta \leq \frac{w^T \tilde{c}}{\sqrt{w^T w}} \tag{3c}$$

where $\xi_i \in \mathbb{R}$ are the slack variables, λ is the regularization hyperparameter, and Eqn.3c is the orientation prior such that $\theta \in (0, 1]$ is the maximum angle that the w is allowed deviate from $\tilde{c}$. Note that we have assumed, without loss of generality, that $||\tilde{c}||_2 = 1$. Fig.7 shows a visualization of this orientation prior.

4.2. Learning

Optimizing Eqn.3 directly seems to be challenging due to the constraint in Eqn.3c. However, it is possible to write the above objective as a conic program. Since conic programs are convex by construction, we can then optimize it using off-the-shelf solvers.

We rewrite Eqn.3c as $\sqrt{w^T w} \leq \frac{w^T \tilde{c}}{\theta}$ and introduce an auxiliary variable $\alpha \in \mathbb{R}$ such that $\sqrt{w^T w} \leq \alpha \leq \frac{w^T \tilde{c}}{\theta}$. Substituting these constraints into Eqn.3 and replacing the SVM regularization term with $\frac{\lambda}{2}\alpha^2$ leads to the conic program:

$$\min_{w,b,\xi,\alpha} \frac{\lambda}{2} \alpha^2 + \sum_{i=1}^{n} \xi_i \tag{4a}$$

$$\text{s.t.} \quad y_i \left(w^T x_i + b \right) \geq 1 - \xi_i, \quad \xi_i \geq 0 \tag{4b}$$

$$\sqrt{w^T w} \leq \alpha \tag{4c}$$

$$\alpha \leq \frac{w^T \tilde{c}}{\theta} \tag{4d}$$

Since the minimum occurs iff $a^2 = w^T w$, Eqn.4 is equivalent to Eqn.3, but in a standard conic form.

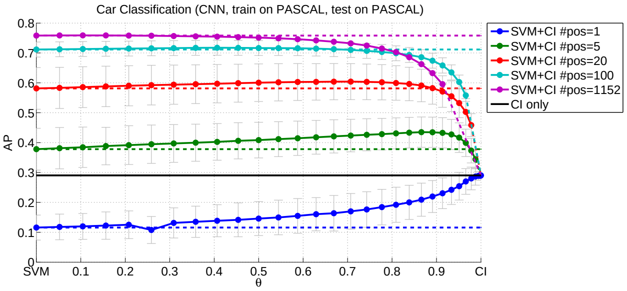


Figure 8: We plot the influence of θ on car classification on PASCAL 2011 with CNN features. When there is only one positive training example available, the imaginary classifier alone ($\theta = 1$) obtains better performance (blue). However, when there is a medium amount of training data (5 to 100) positives, gently incorporating the imaginary classifier ($\theta \approx 0.7$) into an SVM boosts performance (green, red). Error bars show standard deviation over 5 random trials.

We optimize Eqn.4 using the interior point method. In our experiments, we use MOSEK [1]. Optimization took an hour on typical sized problems, but since we use a general purpose solver, improving the implementation will significantly increase run time performance. We note a similar SVM formulation was introduced in [12], but they only impose sign—not orientation—constraints on the weight vector. Moreover, observe that removing Eqn.4d makes it equivalent to the standard SVM.

$\cos^{-1}(\theta)$ specifies the angle of the cone. In our experiments, we found 30° to be reasonable. While this angle is not very restrictive in low dimensions, it becomes much more restrictive as the number of dimensions increases. The probability of a randomly sampled w satisfying the rotation constraint can be determined by calculating the surface area of the spherical cap formed with the cone, then dividing it by the surface area of the whole sphere [25]. The probability of a random w satisfying the angle constraint for 20° in 3D is 0.03, but drops to $O(10^{-48})$ in 100 dimensions.

4.3. Transferring Human Bias into Recognition

Since we believe the bias in the human visual system is favorable, we are interested in transferring this bias into object recognition. To accomplish this, we can train an SVM and impose the imginary classifier as an orientation prior.

Using the same evaluation procedure as the previous section, we compare three approaches: 1) a single SVM trained with only a few positives and the entire negative set, 2) the same SVM with orientation priors on the imaginary classifier, and 3) the imaginary classifier alone. We then follow the same experimental setup as before. We show performance on car classification with CNN features in Fig.8 for varying θ to see the influence of the imaginary classifier on the SVM. Note that with one positive training example (blue curve), the imaginary classifier still provides the best results, suggesting that human bias is more valuable than a single real image.

When we train the standard SVM with five positive examples, the SVM beats imaginary classifiers alone. However, by incorporating an imaginary classifier as an orientation prior into the learning (green curve), the SVM is forced to find a solution that fits the data while agreeing with the human visual system, beating all approaches by nearly 5% AP. These results suggest that transfering the human bias into machine learning methods can improve object recognition performance. Finally, training on the entire dataset (purple curve) gives the best results overall. This is to be expected since large amounts of annotated data is no substitute for noise. However, in the absence of big data, our results suggest extracting knowledge from the human visual system can be powerful.

We show full results for the SVM with orientation priors in Fig.9. In general, imaginary classifiers are able to assist the SVM when the amount of positive training data is only a few examples. In these low data regimes, acquiring classifiers from the human visual system can improve performance by significant margins, sometimes 10% AP.

4.4. Dataset Generalization

Several recent papers have reported that standard computer vision datasets suffer from dataset biases that harm cross dataset generalization performance [35, 30]. Unfortunately, there is no known method to fix it (although there have been several good first attempts [22, 33, 19]). Since the imaginary classifiers are immune to dataset bias (there is no dataset) and instead inherit human biases, our approach can offer some relief.

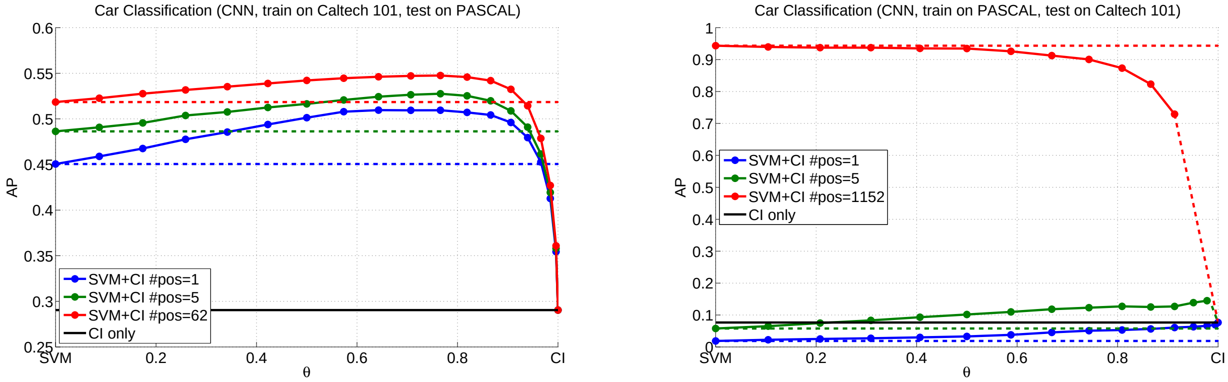
We trained an SVM classifier with CNN features to recognize cars on Caltech 101 [14], but we tested it on object classification with PASCAL VOC 2011. Fig.10a suggest that, by constraining the SVM to be close to the imaginary classifier, we are able to improve the generalization performance of our classifiers, sometimes over 5% AP. We then tried the reverse experiment in Fig.10b: we trained on PASCAL VOC 2011, but tested on Caltech 101. While PASCAL VOC provides a much better sample of the visual world, the orientation priors still help generalization performance when there is little training data available. These results suggest that incorporating the biases from the human visual system can help alleviate some dataset bias issues in computer vision.

5. Bias in the Human Visual System

We have so far examined classifiers acquired from an international population, and our results suggest that there is a

	0 positives		1 positive		5 positives	
Category	Chance	Hum	SVM	SVM+Hum	SVM	SVM+Hum
car	7.3	27.5	11.6	29.0	37.8	43.5
person	32.3	65.6	55.2	69.3	70.1	73.7
tvmonitor	2.6	23.8	38.6	43.1	66.7	68.8
f-hydrant	0.3	5.9	1.7	7.0	50.1	50.1
bottle	4.5	6.0	11.2	11.7	38.1	38.7

Figure 9: We show AP for the SVM with orientation priors for object classification on PASCAL VOC 2011 for varying amount of positive data with CNN features. All results are means over random subsamples of the training sets. SVM+Hum refers to SVM with the imaginary classifier as an orientation prior. Green indicates an improvement of at least 1%.



(a) Train on Caltech 101, Test on PASCAL

(b) Train on PASCAL, Test on Caltech 101

Figure 10: Since the imaginary classifier is estimated only by humans looking at noise, it tends to be biased towards the human visual system, and can alleviate some problems in dataset bias. (a) We train an SVM to classify cars on Caltech 101 that is constrained towards the imaginary classifier, and evaluate it on PASCAL VOC 2011. For every training set size, constraining the SVM to the imaginary classifier with $\theta \approx 0.75$ is able to improve generalization performance. (b) We train a constrained SVM on PASCAL VOC 2011 and test on Caltech 101. For low data regimes, the imaginary classifier is able to again boost performance.

bias from the human visual system that influences the mental images that people imagine. However, everyone does not necessarily share the same bias with each other.

We found that people from India and the United States may have different mental images of sports balls. We instructed workers on Mechanical Turk to find "sport balls" in CNN noise, and clustered workers by their geographic location. Fig.11 shows the imaginary classifiers for both India and the United States. Even though both sets of workers were labeling noise from the same distribution, Indian workers seemed to imagine red balls, while American workers tended to imagine orange/brown balls. Remarkably, the most popular sport in India is cricket, which is played with a red ball, and popular sports in the United States are American football and basketball, which are played with brown/orange balls. We hypothesize that Americans and Indians may have different mental images of sports balls in their head and the color is influenced by popular sports in their country. This effect is likely attributed to phenomena in social psychology where human perception can be influenced by culture [6, 4]. Since environment plays a role in the development of the human vision system, people from different cultures likely develop slightly different images inside their head.

This effect can be observed on more categories, sometimes manifesting in subtle ways. We created a classifier for each country, but this time asked workers to find cars in CNN noise. Fig.12 shows the distribution of top poses that each car imaginary classifier finds. Surprisingly, the American imaginary classifier favors left-right facing cars, while the Singaporeans favor front-back views of cars. This result suggests that people may different biases in their human visual system.

6. Discussion

While the ideas in this paper may seem unconventional, they highlight how human imagination can be a rich resource for computer vision systems. Humans are able to imagine objects under any transformation, even for concepts never before seen. However, creating intelligent vision machines with the capability to imagine radically new concepts never before encountered in its data remains a significant, open research problem in our field. This paper explores this direction by showing that it is possible to transfer

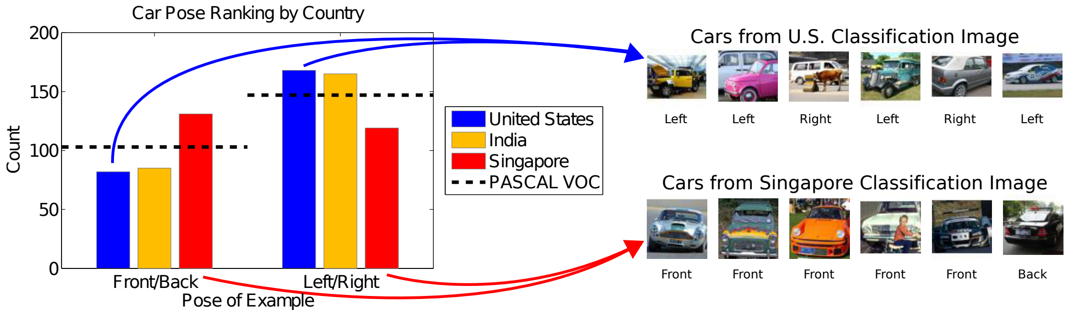
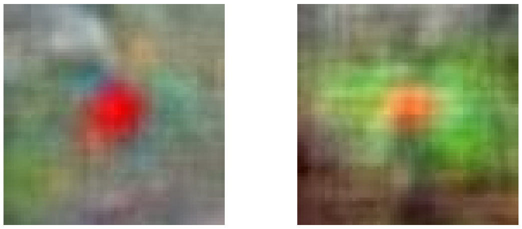


Figure 12: We created an imaginary classifier for each country, and examine the distribution of poses that each classification image favors. Notice that US workers seem more likely to imagine left/right facing cars, while Singapore workers may favor imagining front/back facing cars. Please see text for details.



(a) India

(b) United States

Figure 11: By instructing workers to classify CNN noise as a sports ball or not, then creating imaginary classifiers by country (shown above), we reveal the different mental images of sports ball (the red/orange circles in the center) that people from different countries have inside their head. Indians seem to imagine a red ball, which is the standard color for a cricket ball and the predominant sport in India. Americans seem to imagine a brown or orange ball, which could be an American football or basketball, both popular sports in the U.S.

classifers extracted from human imagination into a machine with modest success. Our hope is that our ideas will inspire future work on building machines with the ability to imagine new visual concepts just like a human.

Acknowledgements: We thank Aditya Khosla for inspiring discussions, and Andrew Owens and Zoya Bylinskii for helpful comments. Funding was provided by a NSF GRFP to CV, and a Google research award and ONR MURI N000141010933 to AT.

References

[1] The MOSEK Optimization Software. http://mosek.com/. 7

[2] A. Ahumada Jr. Perceptual classification images from vernier acuity masked by noise. 1996. 1, 2

[3] B. L. Beard and A. J. Ahumada Jr. A technique to extract relevant image features for visual tasks. In *SPIE*, 1998. 2

[4] C. Blais, R. E. Jack, C. Scheepers, D. Fiset, and R. Caldara. Culture shapes how we look at faces. *PLoS One*, 2008. 8

[5] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie. Visual recognition with humans in the loop. 2010. 2

[6] H. F. Chua, J. E. Boland, and R. E. Nisbett. Cultural variation in eye movements during scene perception. *Proceedings of the National Academy of Sciences of the United States of America*, 2005. 8

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 1, 4

[8] E. d'Angelo, A. Alahi, and P. Vandergheynst. Beyond bits: Reconstructing images from local binary descriptors. *ICPR*, 2012. 3

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 4, 5

[10] J. Deng, J. Krause, and L. Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. *CVPR*, 2013. 2

[11] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv*, 2013. 4

[12] A. Epshteyn and G. DeJong. Rotational prior knowledge for svms. In *ECML*. 2005. 2, 7

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge. *IJCV*, 2010. 4

[14] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *PAMI*, 2006. 7

[15] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 2010. 4

[16] M. Ferecatu and D. Geman. A statistical framework for image category search from a mental picture. *PAMI*, 2009. 2

[17] F. Gosselin and P. G. Schyns. Superstitious perceptions reveal properties of internal representations. *Psychological Science*, 2003. 2

[18] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classification. *ECCV*, 2012. 3

[19] J. Hoffman, E. Tzeng, J. Donahue, Y. Jia, K. Saenko, and T. Darrell. One-shot adaptation of supervised deep convolutional models. *arXiv*, 2013. 7

[20] A. A. Jr and J. Lovell. Stimulus features in signal detection. *The Journal of the Acoustical Society of America*, 1971. 2

[21] H. Kato and T. Harada. Image reconstruction from bag-of-visual-words. In *CVPR*, 2014. 3

[22] A. Khosla, T. Zhou, T. Malisiewicz, A. Efros, and A. Torralba. Undoing the damage of dataset bias. *ECCV*, 2012. 7

[23] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1, 2, 4

[24] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1989. 2

[25] S. Li. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics and Statistics*, 2011. 7

[26] M. C. Mangini and I. Biederman. Making the ineffable explicit: Estimating the information employed for face classifications. *Cognitive Science*, 2004. 2

[27] R. F. Murray, P. J. Bennett, and A. B. Sekuler. Optimal methods for calculating classification images: Weighted sums. *Journal of Vision*, 2002. 3

[28] P. Neri. Estimation of nonlinear psychophysical kernels. *Journal of Vision*, 2004. 3

[29] D. Parikh and C. Zitnick. Human-debugging of machines. In *NIPS WCSSWC*, 2011. 2

[30] J. Ponce, T. L. Berg, M. Everingham, D. A. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B. C. Russell, A. Torralba, et al. Dataset issues in object recognition. In *Toward category-level object recognition*. 2006. 7

[31] R. Salakhutdinov, A. Torralba, and J. Tenenbaum. Learning to share visual appearance for multiclass object detection. In *CVPR*, 2011. 2

[32] A. Sorokin and D. Forsyth. Utility data annotation with amazon mechanical turk. In *CVPR Workshops*, 2008. 2, 3

[33] T. Tommasi, N. Quadrianto, B. Caputo, and C. H. Lampert. Beyond dataset bias: multi-task unaligned shared knowledge transfer. 2013. 7

[34] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *JMLR*, 2002. 2

[35] A. Torralba and A. Efros. Unbiased look at dataset bias. In *CVPR*, 2011. 3, 7

[36] S. Vijayanarasimhan and K. Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. In *CVPR*, 2011. 2

[37] L. Von Ahn, R. Liu, and M. Blum. Peekaboom: a game for locating objects in images. In *SIGCHI Human Factors*, 2006. 2

[38] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba. HOGgles: Visualizing Object Detection Features. *ICCV*, 2013. 1, 3, 4

[39] P. Weinzaepfel, H. Jégou, and P. Pérez. Reconstructing an image from its local descriptors. In *CVPR*, 2011. 3

[40] P. Welinder, S. Branson, P. Perona, and S. J. Belongie. The multidimensional wisdom of crowds. In *NIPS*, 2010. 2